Capstone Project - Predicting User Sentiments

By Sarafudheen M Tharayil

Business Problem

You are hired as a data scientist by a restaurant chain (probably the most prestigious) to tune the business based on customer review. It is found that recently there is a decline of 5% in the revienue in the last quarter. The restaurant management want to analyze the current situation and see what the individual customer feel about the restaurant.

The Restaurant management decided to give a bonus point for their customer if they give a review (good or bad) in foursquare website in their venue in the chain of business.

As there are many outlets and many thousands of users for the restaurants looking at each comment and review is not possible. So, as a data scientist, you are asked to prepare a model to predict if the user is happy or unhappy based on the comments without reading each and every line in it.

The management also want to see what are the main keyword in the review (positive and negative) based on which the management can take actions.

Overview of the solution

A robust prediction model need to be prepared based on the available reviews and scroes that is available. The available source for sentiments and its rating need to be reliable. Once we recive the data(in our case we got a **keggle dataset**), the data need to be analyzed and make sure that we have enough data to try with.

Once we get the data, we do cleaning, remove stop words and vectorize it so that we can apply different machine learning algorithms. In our case, we use a classification algorithm to classify the comments from 1 - 3.

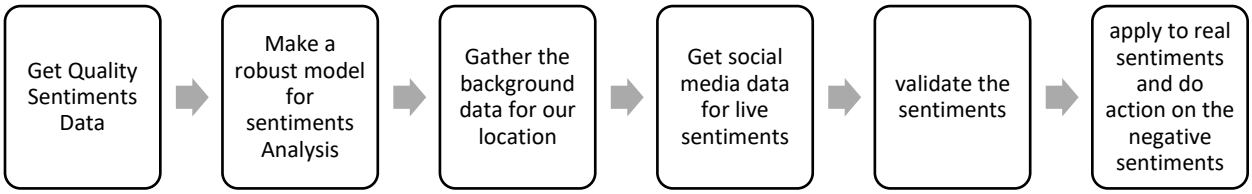
Based on the words, we can find different words which can be identified as bad reviws and make some special care about those cases.

Once we prepare the model, we could use the model to find the specific venue in the foursqure and find the recent user comments. Also check if the user comments are positive or negative. If they are negative, take special care and make necessary actions to rectify those bad reviews based on the keywords used in the review.

We use sentiment analysis in sci-learn to make a predictive model in order to see if the user review is positive, negative, or neutral.

We will use the available sentiments from different restaurants to train the model. Based on the trained model, we will test the prediction based on sample test data. once the testing is found successful, we will use the new reviews from the user and give a score based on the trained model

The process can be shown as in the following picture:

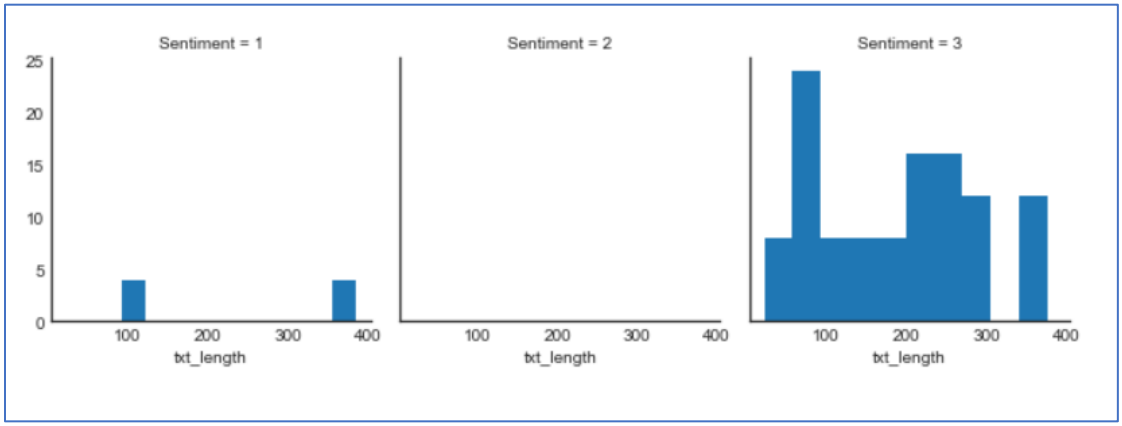


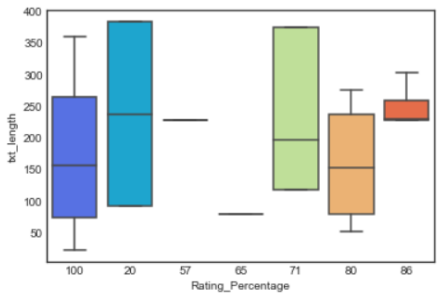
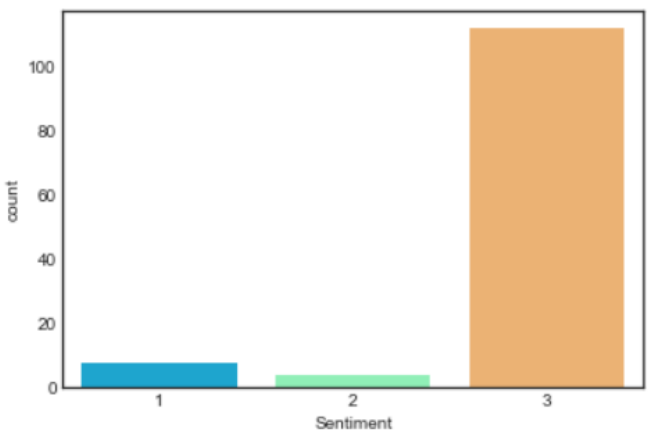
Data Preparation, Scarping, Cleanup

Initially we thought of getting the data from Foursquare, but it is found that the data for tips is premium category, so it has a limit to get the number of reviews. We have searched for different datasets for Indian context and found the data found a good one at https://www.kaggle.com/ranjitha1/hotel-reviews-city-chennai/version/2. This dataset is good for experimentation purpuse. We use this dataset and downloaded to 'chennai_reviews.csv' and we read data to dataframe

We completed different inferential statistics. Some of them are displayed below:

Use FacetGrid from the seaborn library to create histograms of text length based off of the ratings:



	
Creating a boxplot of text length for rating	Creating a count plot of the number of occurrences

Sentiments Analysis - Classification

The classification algorithms need numerical feature vector in order to perform the classification task.In this section we'll convert the raw messages (sequence of characters) into vectors (sequences of numbers). We applied Text Normalization with Bag-of-Words strategy and takes care of the words processing. It transforms the entire data frame with a reduced memory space etc. Then vectorization convert messages to vector. We convert each message, represented as a list of tokens having frequency of word occurrences, weight of the counts, normalize the vectors to unit length

Using TF-IDF & Nave-Bayes For Sentiments Analysis¶

TF-IDF is a weight often used in information retrieval and text mining. This weight is a statistical measure used to evaluate how important a word is to a document in a collection or corpus. Using scikit-learn Naive Bayes classifier in our scenario. The data is fit into the MultinomialNB class object. Model Evaluation is done using the confusion matrix. As there are three values in the set, it is found all the values are correctly predicted without any false positives.

Applying the Sentimental Model in Real Data

Once we model the data and once he data is gathered from the FOURSQUARE website, we applied the logic to use the model in the actual scenario of the hotel we are interested in. We utilize the fourquare APIs to get the sentiments of the users and see if they are in favor of the restaurant or not We gather details on the hotel data¶ based on cities to cross check cities of "UAE" data.

Foursquare API for live sentiments

We get the details of the cities and interested location based on the query that we set. and then based on the interested location, we will gather the sentiments of specific location.

We first gather all the interesting venues with the food type.

- We then save all the interesting places for analysis
- then we select the required interested venue for focused study

The information about the venues are gathered for further analysis.

Gather Sentiments & Prediction ¶

We gather sentiments for the interested venues. At this time we utilize the sentiments analytics model we created as a classification model based on Nave Bayes. We used the trained model based on the good quality dataset we got before. Optionally, we can gather more similar comments from Foursquare and update the training database

Result Achieved

The following diagram shows classification report based on SK-LEARN.

	precision	recall	f1-score	support
1	0.33	1.00	0.50	1
2	1.00	1.00	1.00	1
3	1.00	0.91	0.95	23
avg / total	0.97	0.92	0.94	25

Result with training data

	precision	recall	f1-score	support
3	1.00	1.00	1.00	1
avg / total	1.00	1.00	1.00	1

Result with live sentiments

We have a F1-Score of (.95), which is very good for the current data set. This can be further improved.

Observations, Recommendations & Final Remarks¶

The score says the quality of the model is good enough for this scenario. Further improvements to the model can be done based on more refined data and trying different algorithms with different hyperparameter.

It is observed that FOURSQUIRE API cannot bring the data for the sentiment's analysis. For more accurate predictions, we need live data which need paid account. As this is not as part of

our exercise, we will leave to the reader to go for further refinement of the model based on more live data and retrain.

Conclusion

It is exciting to experience the data science work and as we enjoy the work. The curiosity of getting the result, in our case the result of showing the sentiments of customer will lead us to make the results. This project will pave a foundation for our career.